Optimization of Data Mining for Grouping Courses Using the MDDS and MAR Methods

Paryati[1(✉)], Agung Mulyo Widodo[2], Shankar Rao Munjam[3], Krit Salahddine[4], and Sagayam Martin[5]

[1] Informatic Engineering Faculty Technic, UPN "Veteran" Yogyakarta, Yogyakarta, Indonesia
yaya_upn_cute@yahoo.com
[2] Computer Science, Esa Unggul University, Jakarta, Indonesia
agung.mulyo@esaunggul.ac.id
[3] Mathematic and Computer Science, Woxsen University, Hyderabad 502345, India
Shankar.rao@woxsen.edu.in
[4] Compter Science, Ibn Zohr University, Quzazate City, Morocco
[5] Electrict Engineering, Karunya University, Lahore City, India

Abstract. To find meaningful clusters from a data set, attribute clustering is carried out, so that the attributes in the created cluster will have a high or very good correlation, as well as interdependence with each other. While the attributes in the other clusters are less correlated or more independent. The experimental results show how to determine the list of dominant attribute ratings using soft set theory. A series of experiments were conducted to evaluate the clustering performance, clustering efficiency and scalability of the MAR and MDDS algorithms. The experimental results show that, MDDS achieves better clustering accuracy and stability than the MAR algorithm, at the same time increasing efficiency. MDDS has clear advantages over MAR on large data sets in terms of clustering efficiency as well as clustering accuracy. In addition, the MDDS technique has better scalability. It can be applied to small category data sets as well as large category data sets. The clustering of data under soft set theory can be considered as a technique for data mining. Maximum Degree of Domination in soft set theory is applied to select grouping attributes. In the assessment of student lectures to determine the optimal clustering attributes, and get the best value is very urgent in data clustering. So in the assessment of lecturers' lectures on the subjects being taught, in order to get optimal results, clustering of lecture assessments is needed. Actually, there are five types of methods and techniques, based on coarse and soft sets, to select the attributes for grouping course assessments, namely TR, MMR, MDA, NSS, and MAR. However, the MAR method has better numerical computational time, compared to the other four approaches. In the MAR method, there is a drawback, namely the execution time is still slow, because in the iteration process it determines the relative attributes. So to overcome these problems, use an alternative technique based on soft sets to select clustering attributes, namely the Maximum Degree of Domination in Soft set theory (MDDS) method. In this method, the steps in defining the multi soft set are explained first. Then determine the dominance of the soft set and its degree. Then the maximum degree of dominance will be used to determine the best grouping attributes in the assessment of student lectures. The results of the experimental data obtained show that the MDDS technique is very good, and can significantly reduce the numerical computation time.

I G. P. Suta Wijaya et al. (Eds.): MIMSE-I-C 2022, ACSR 102, pp. 169–183, 2022.
https://doi.org/10.2991/978-94-6463-084-8_16

The MDDS method is better than MAR with a working percentage of 43.99%. The MDDS method also has better scalability, which is indicated by the execution time increasing linearly, with increasing data size. While the accuracy of the experimental data set has a class attribute, and has increased by 3.23%. So the MDDS technique can be a solution to the problem solving above, so that in the assessment of lecturers' lectures on the subjects being taught, they can get optimal results.

Keywords: Optimization · Data Mining · MDDS · MAR

1 Introduction

1.1 Background

Many clustering techniques are used to partition large data sets. Several clustering techniques were based on rough set theory and soft set theory. However, existing techniques have various limitations, namely a long execution time and low accuracy. Evaluation of clustering technique for categorical data is a difficult job. Currently, the most widely used criteria for evaluation includes accuracy and efficiency of clustering. Clustering accuracy measures the quality of the clustering technique. A higher value of the clustering indicates better clustering results. Clustering efficiency is measured by the running time of the technique. The length of execution time, the low efficiency (Deng et al. 2012) have shown that the clustering technique has low accuracy, while having a higher efficiency compared to direct optimization-based techniques. On the contrary, direct optimization-based techniques have high accuracy, while the grouping has a low efficiency compared with a heuristic technique. Therefore, a new clustering technique for categorical data with high efficiency and high classification accuracy is required.

1.2 Objectives and Scope

This objective: To study and develop an attribute-oriented clustering technique on the data categories that improves efficiency and higher accuracy. And to apply the proposed technique in educational data mining. The scope of this research is to study and analyze several clustering attribute selection techniques for categorical data. Based on the disadvantages of previous techniques, it should be proposed a new technique be introduced that has better performance. The execution time and accuracy of the proposed technique will be compared with previous technique in seventeen UCI benchmark datasets machine learning. Finally, the new technique will be applied in educational data mining.

2 Literature Review

2.1 Knowledge Discovery in Database

Knowledge Discovery in Databases (KDD) and data mining have been attracting a significant amount of attention from research, industries, and media of late. Data mining

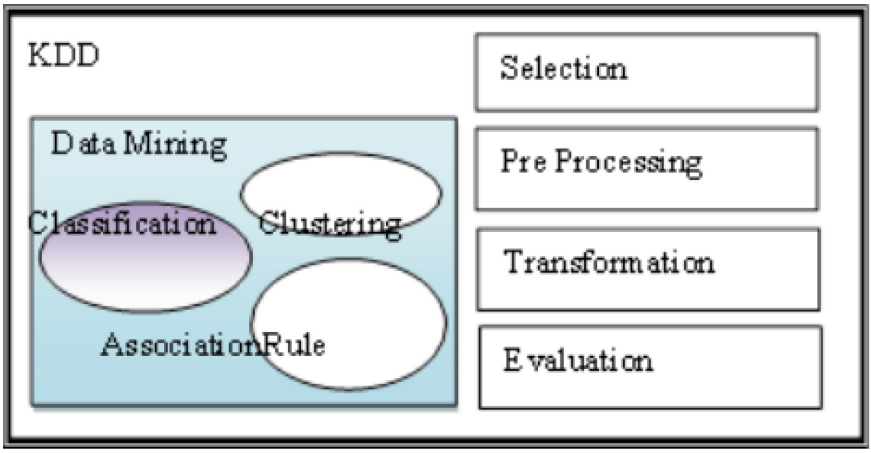

Fig. 1. Relation between KDD, data mining, and clustering

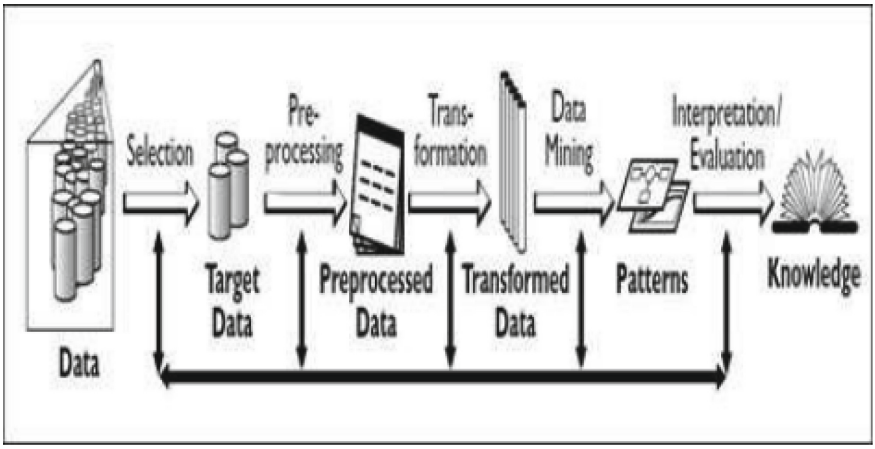

Fig. 2. Overview of the steps constituting the KDD process

is part of KDD. While one branch of data mining is the clustering. The relationship between KDD, data mining, and clustering can be shown in Fig. 1.

KDD is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. KDD is a multi-step process that turns raw data into useful knowledge (Bagga and Singh 2011). KDD can also be defined as the entire process of converting raw data into useful knowledge which consists of a series of phase transformations, such as data pre-processing and postprocessing. Briefly KDD is the process of discovering which is deemed as useful knowledge from the data set. Knowledge Discovery in Databases brings together current tree search on the exciting problem of discovering useful and interesting knowledge in databases. While data mining is one of the overall process of the KDD process. The KDD process is shown in Fig. 2.

2.2 Data Mining

Data mining is referred to the process of extracting hidden and useful information in large data repositories. The existence of data mining with the data explosion problem (bigdata) has recently occurred in many organizations (purchasing data, sales data, customer data, transaction data, etc.). Almost all the data was entered using a computer application used to handle the day-to-day transactions that are mostly on-line Transaction Processing (OLTP). Imagine how many transactions had been entered into by a Care four hypermarket or some sort of credit card transactions from a bank in a day and imagine how big the size of their data, if it has been running a few years later. The data warehouse supports On Line Analytical Processing (OLAP), the functional and performance requirements of which are quite different from those of the on-line transaction processing applications traditionally supported by the operational databases. Data warehouses provide OLAP tools for the interactive analysis of multidimensional data of varied granularities, which facilitates effective data mining.

2.3 Classification

Classification is one of the most common learning models in data mining (Ahmed 2004). The goal is to build a model to predict the behavior of something through the database records which are classified into a number of standard classes based on certain criteria (Ahmed 2004). Equipment used for classification are neural networks, decision trees, and if-then-else rules. Classification process is usually divided into two phases: learning and test. In the learning phase, most of the data that has been known to form a class of data is fed to the model estimates. Test phase models are tested with most other data to determine its accuracy. With sufficient accuracy this model can be used to predict the class of data that is not known yet.

2.4 Clustering

Data clustering is one of the basic tools available, to understand the structure of the data set. The process of grouping a set of physical or abstract objects into classes of similar objects is known as clustering. Clustering techniques play an important role in machine learning, data mining, information retrieval, web analytics, marketing, medical diagnostics, and pattern recognition. Clustering is often called unsupervised learning task because there is no class that shows the value of a prior clustering given from the data sample, which is the case in supervised learning. General definition of clustering could be “the process of organizing objects into groups whose members are similar in some ways”. Therefore, the cluster is a collection of data objects that are similar to each other in a same and distinct cluster with objects in other clusters. Clustering is a dynamic area of research on data mining. Many clustering algorithms are constantly being developed. The selection algorithm clustering depends on the types of data available and the purpose of an application.

2.5 The Basic Steps of the Clustering Process

The process of clustering may lead to different partitions of a data set, depending on the specific criteria used for clustering. Thus, there is a need of preprocessing before the user assumes the task of grouping a set of data. The basic steps for developing a clustering process can be presented as follows:

1. Feature selection. The goal is to select properly the features on which clustering is to be performed so as to encode as much information as possible concerning the task of interest. Thus, preprocessing of data may be necessary prior to their utilization in clustering task.
2. Clustering algorithm. This step refers to the choice of an algorithm that results in the definition of a good clustering scheme for a data set. Proximity measure and a clustering criterion mainly characterize a clustering algorithm as well as its efficiency to define a clustering scheme that fits the data set.

2.6 Maximum Attribute Relative (MAR) Technique

Here max refers to the value that is the highest in the probability distribution and mode refers to the value that is most frequently occurred in the probability distribution. The details of the MAR algorithm is shown in Fig. 3.

The MAR technique uses complex mathematical models, because this technique calculates the value of support, max_support and min_support. Support value = 1 is summed as max_support, while others as min_support. Clustering attribute is determined based on the largest amount of max_support. If there are greatest max_support of more than 1, then the clustering attribute is an attribute that has the largest value of min_support. From the analysis of the limitations of the MAR technique, there is a need

```
Input: Categorical-
valued data-setOutput:
A Clustering attribute
Begin
1. Builds the multi-soft set approximation
2. Calculate Support, MaxSup
    and MinSuPfor i = all
    categories
        for j = all categories
            intersection = Data(i)
            And Data(j)Sup(i,j) =
            Intersection/Data(j)
            if Sup (i,j) = 1 then
```

Fig. 3. The MAR algorithm

to develop technique for clustering. This new technique is called the Maximum Degree of Domination in Soft set theory (MDDS).

3 Method

3.1 Maximum Attribute Relative (MAR) Technique

The MAR technique (Mamat et al. 2013) a pair (F, A), refers to multi-soft sets over the universe U representing a categorical value information systems S = (U,A,V,f).

3.2 Maximum Degree of Domination in Soft Set Theory (MDDS) Technique

The proposed MDDS, for selecting a clustering attribute is as follows. First, the presentation of the idea of multi-soft sets is to deal with multi-valued information system. Second, the presentation of the notion domination in multi-soft sets. Finally, MDDS is presented to select the best clustering attribute.

3.3 The MDDS Technique

Let (F, A) be multi-soft set over U representing S = (U,A,V,f), based on, the soft set (F, ai) with maximum degree of domination will be selected as a clustering i.e. max{k1, k2,…, kn}.

3.4 Complexity of MDDS

Just as in the MAR technique, suppose that in an information system, there are n objects, m attributes and l is the maximum distinct values of each attribute. The computational cost to determine the elementary set of all attributes is nm. The proposed technique needs m(m − 1) times to determine the support for each category. The computational complexity for the proposed technique is O(nm + m(m − 1)). After compared with MAR technique, it is clear that the proposed technique has a lower complexity.

4 Experimental Result

4.1 Experimental Design

This section, the discussion of upon the comparison between MDDS and MAR. While the main focus of the experiments is on the performance measurement of the proposed technique in which execution time and accuracy are used as a parameter.

4.2 Data Sets

For comparisons, two techniques which have been discussed will be used with seventeen datasets obtained from the benchmark UCI machine learning repository and a supplier dataset.

4.3 Language and Platform for Implementation

The two techniques MAR and MDDS are implemented using Matlab programming language version R2009a under Windows 13 Home Edition operating system powered by Intel i7 processor with 16 GB memory.

4.4 Performance Analysis

Validating the clustering results is heavy work. It needs to be measured against the accuracy results clustering in a certain way. The methods are used to evaluate the measures of the accuracy of clustering, in addition to the accuracy, the computing time is also very important as it relates to efficiency. The faster the results obtained, decision making can be done faster.

4.5 Execution Time

In this sub-section, the experimental results of the two techniques will be presented. The execution time in selecting a clustering attribute is presented in Table 1.

4.6 The Results of the Comparison Are Made Between the MDDS and MAR in Terms of Execution Time

To calculate the increase in the relative improvement MAR and MDDS the following formula is used (Fig. 4):

$$(\%) = \frac{|MAR - MDDS| \times 100\%}{MAR} \tag{1}$$

Table 1. Execution time of comparison results.

No	Data set	Atr	Inst	Clas	Decision		Ac		Improve %
					Mr	Md	Mar	Mds	
1	Balanc	4	63	3	1	1	0.64	0.64	0.0
2	Car	6	18	4	4	1	0.70	0.70	0.0
3	Lenses	4	24	3	2	1	0.63	0.63	0.0
4	Cancer	56	32	3	1	47	0.69	0.72	3.4
5	Monk	6	43	2	3	1	0.50	0.50	0.0
6	Mushr	21	80	2	2	15	0.52	0.52	0.0
7	Nurse	8	13	5	6	1	0.34	0.42	23
8	Solar F	10	2	3	9	10	0.91	0.91	0.0
Average of overall improvement 3.23									

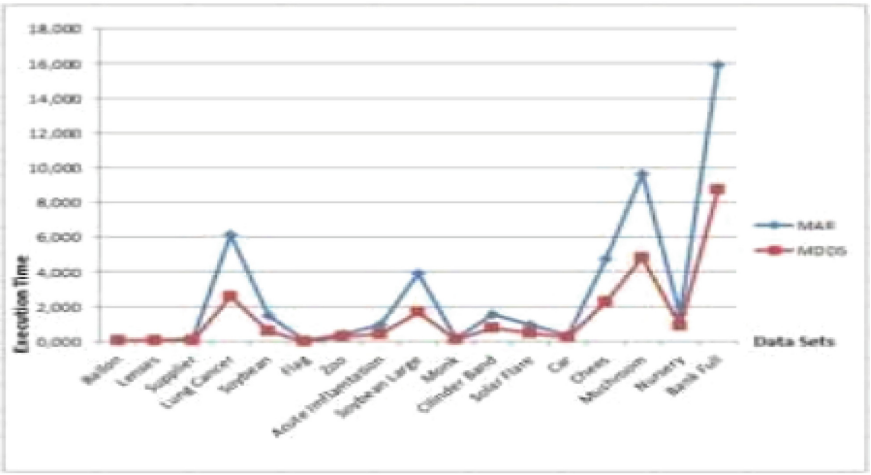

Fig. 4. The scalability of MAR and MDDS to the number of instances

4.7 Scalability of MDDS to the Size of Data

In reviewing the scalability of MAR and MDDS techniques on the seventeen data sets, the value varies in several numbers of instances and attributes. However, on the whole, these techniques have a good scalability to data size. The data size here is based-on the complexity of computing entries in data table. From the seventeen data sets in Table 2 Fig. 5 described the scalability of these techniques to the data size. It is clearly depicted their good scalability to the data size.

5 Results of Implementation of Lecture Assessment Data Set

5.1 Data Sets of Assessment

This section explains and discusses the experimental results of the proposed technique. The main focus of the experiments is on the performance measurement of the proposed technique in which execution time is used as a parameter. The Data is taken from the evaluation of Information Engineering and Architecture Department. The data is shown in Table 3.

All the data we taken from University Pembangunan National "Veteran" Yogyakarta country Indonesia for four years with periods from 2018 to 2021. The assessment consists of several attributes which were different. Each of the majors and courses does not have the same assessment criteria, all of it are in the form of assignments, midterm and final exams. Midterms be done in the middle of the semester and is done in writing. The final exams are given at the end of the semester. Both are done on a scheduled basis. Students' name, age, race, and finance were not necessary in this assessment (Table 4).

Table 2. Accuracy comparison results.

No	Data Sets	MAR	MDDS	Improvement %
1	Acute Imflam	0.958	0.394	58.87
2	Ballon	0.051	0.023	54.90
3	Bank Full	31.824	17.472	45.10
4	Car	0.327	0.246	24.77
5	Chees	4.763	2.274	52.26
6	Cylinder Band	1.560	0.780	50.00
7	Flag	0.020	0.015	25.00
8	Lenses	0.038	0.027	28.95
9	Lung Cancer	6.158	2.597	57.83
10	Monk	0.146	0.101	30.82
11	Mushroom	9.610	4.824	49.80
12	Nursery	1.552	0.877	43.49
13	Solar Flare	0.931	0.453	51.34
14	Soybean	1.477	0.596	59.65
15	SoybeanLarge	3.885	1.670	57.01
16	Supplier	0.173	0.056	67.63
17	Zoo	0.399	0.262	34.34
Average of overall improvement 43.99				

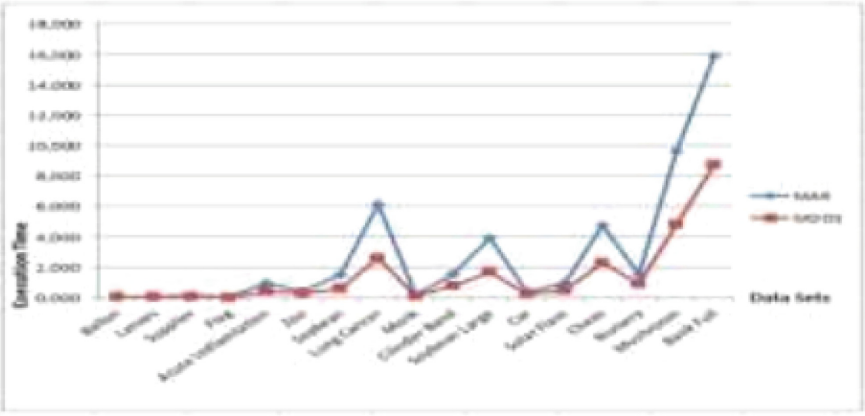

Fig. 5. The scalability of MAR and MDDS to the number data size

Table 3. Data sets of Assessment

Attribute	Domination Degree					Max Domination
	T1	T2	T3	MT	FE	
T1	0.00	0.00	0.01	0.01	0.00	0.01
T2	0.01	0.00	0.02	0.01	0.00	0.02
T3	0.01	0.05	0.00	0.00	0.00	0.05
MT	0.01	0.38	0.02	0.00	0.00	0.38
FE	0.01	0.00	0.01	0.01	0.00	0.01

Table 4. Transformation data assessment into categorical data

No	Data Assessment	Category
1	0–20	1
2	21–40	2
3	41–60	3
4	61–80	4
5	81–100	5

Table 5. Assessment of Artificial Intelligence course

No	Courses	Object	Attribute
1	Algorithm	99	5
2	S Engineering	260	6
3	System Security	269	6
4	File System	190	5
5	Artificial Intelligence	173	5
6	Database	88	5
7	Architecture Design	94	10
8	Design Studio	34	29
9	Architecture Studio	181	7
10	Contextual Studio	34	30

5.2 Data Descriptions

1. Artificial Intelligence Course
   Assessment of Artificial Intelligence course taken in 2021 has five attributes, namely task 1 to task 3, mid-term, and final exam as shown in Table 5.

Table 6. Matrix results from Software Engineering course

Attribute (with respect to)	Domination Degree						Maximum Domination
	T1	T2	T3	T4	MT	FE	
T1	0.00	0.00	0.00	0.00	0.00	0.00	0.00
T2	0.01	0.00	0.00	0.00	0.00	0.00	0.01
T3	0.01	0.00	0.00	0.00	0.00	0.00	0.01
T4	0.01	0.00	0.00	0.00	0.02	0.00	0.02
MT	0.01	0.03	0.00	0.03	0.00	0.00	0.03
FE	0.01	0.00	0.00	0.00	0.02	0.00	0.02

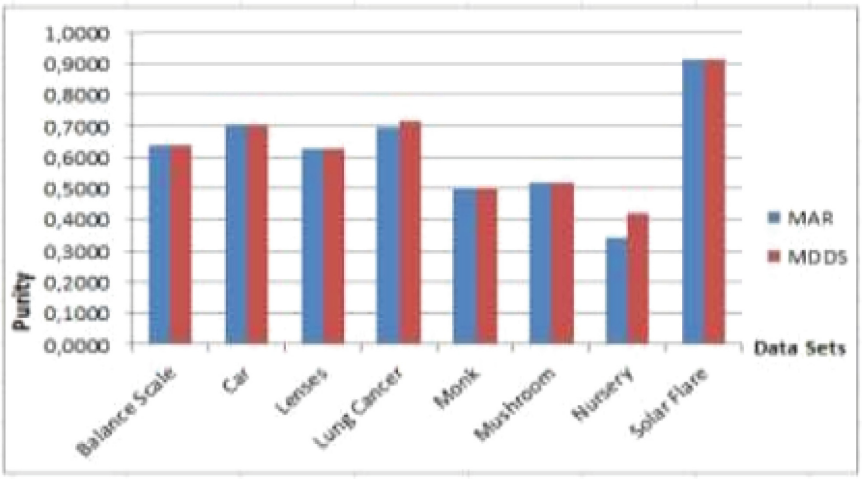

Fig. 6. The accuracy of MAR and MDDS

Table 6 is a matrix that indicates the degree of dominance attribute value of Artificial Intelligence course in 2021. The MDDS technique provides a MT as the most dominant attribute compared to other attributes, where the maximum domination is 0.38. Hence MT is selected as a clustering attribute, whereas visualization is divided in three clusters as shown in Fig. 6 (Fig. 7).

Table 6 is a matrix that indicates the degree of dominance attribute value of Software Engineering course.

MDDS technique provides a MT as the most dominant attribute compared to other attributes, where the maximum domination is 0.03. Hence MT is selected as a clustering attribute, whereas visualization is divided in five clusters as shown in Fig. 8.

Table 7 is a matrix that indicates the degree of dominance attribute value of Databases course in 2021. The MDDS technique provides a T3 as the most dominant attribute compared to other attributes, where the maximum domination is 0.1. Hence T3 is selected

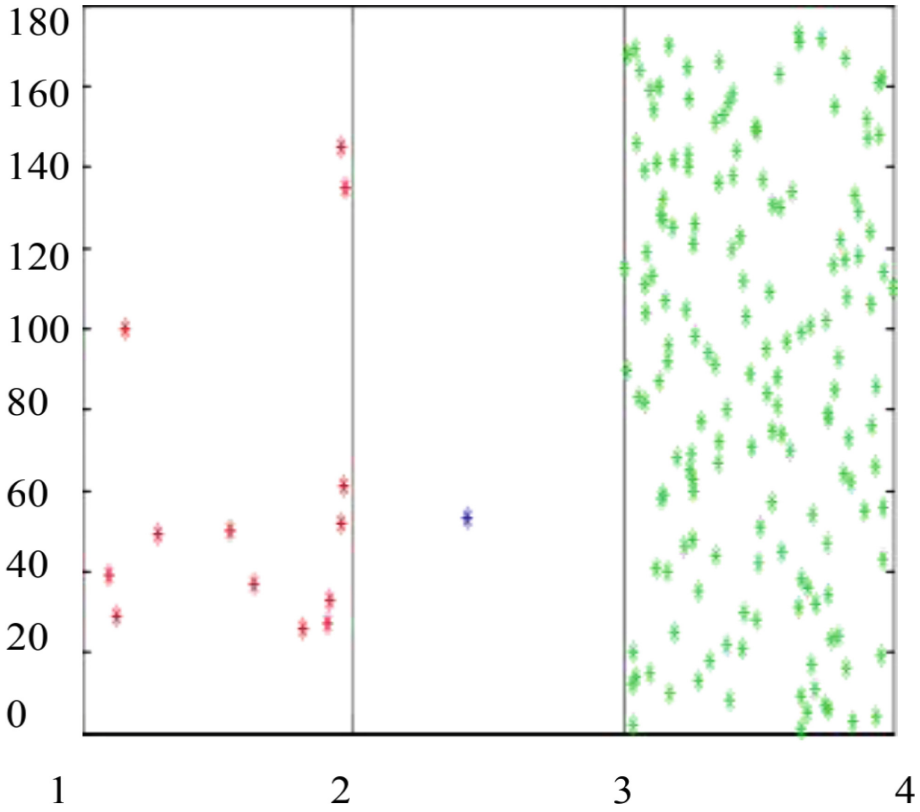

Fig. 7. Clustering visualization of students on Artificial Intelligence

as a clustering attribute, whereas visualization is divided in three clusters as shown in Fig. 9.

6 Conclusion

A series of experiments were conducted to evaluate the clustering performance, clustering efficiency and scalability of MAR and MDDS algorithms. The experimental result show that MDDS achieves better clustering accuracy and stability than MAR algorithm, at the same time increases the efficiency. MDDS has obvious advantage against MAR on large data sets in terms of clustering efficiency as well as clustering accuracy. In addition, The MDDS technique has better scalability. It can be applied on small categorical data sets as well as large categorical data sets.

From the analysis of the limitations of the MAR technique, there is a need to develop clustering algorithm for data categories. The proposed technique is the Maximum Degree of Domination in Soft Set theory (MDDS). The steps of MDDS technique are as follows:

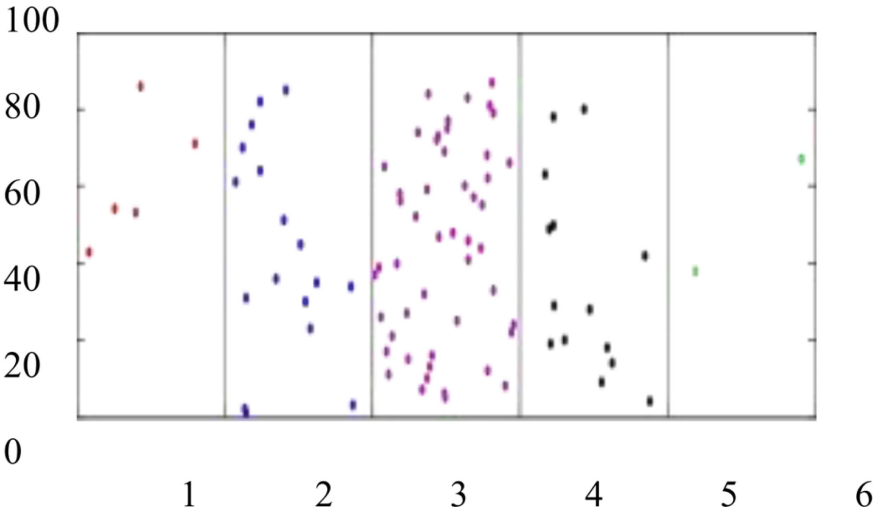

Fig. 8. Clustering visualization of students on Software Engineering course

Table 7. Matrix results from Databases course

Attribute (with respect to)	Domination Degree					Maximum Domination
	T1	T2	T3	MT	FE	
T1	0.00	0.00	0.08	0.02	0.00	0.08
T2	0.01	0.00	0.01	0.00	0.00	0.01
T3	0.10	0.00	0.00	0.02	0.02	0.10
MT	0.10	0.00	0.01	0.00	0.02	0.10
FE	0.00	0.00	0.01	0.00	0.00	0.01

- Build the multi-soft set approximation.
- Calculate Domination of Attributes.
- Select maximum of domination degree of each attribute.
- Select clustering attribute based max degree domination.

As the input is categorical data and the output is the clustering attribute, MDDS can overcome the limitations of MAR. The MDDS technique has succeeded in improving performance. The execution time and the number of iterations is lower than the MAR technique. Average execution time of 17 data sets is 43.99% faster, at the same time the average number of iterations of the 17 sets of data is reduced to 15.26%. While the accuracy of eight data sets which have a class attribute has increased 3.23%. The number of clusters was not determined from the start, so it will be more for user convenience.

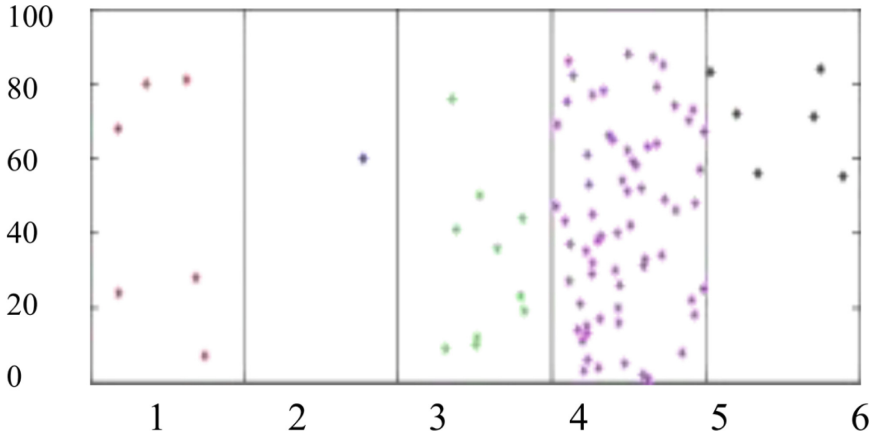

Fig. 9. Clustering visualization of students on Databases course

Suppose that in an information system, there are n objects, m attributes and l is the maximum distinct values of each attribute. Computational cost to determine the elementary set of all attributes is nm.

The MAR and MDDS techniques were applied to select attributes of clustering on the data sets assessment. Results of this experiment showed that the selection of a dominant attribute of the data sets assessment can be performed faster than the MAR technique. The speed increases by up to 50.49%. This speed is obtained because of the simplification process, so that the number of iterations is reduced. As the selection process attributes can be done faster, so the clustering of students will also be faster. But accuracy can't be determined because the dataset there has no decision attribute. The MDDS gives better results than the previous techniques, however it also has some limitations, which includes: The MDDS is more focused on categorical data, whereas in a real database, variety and range of data is enormous. Data is transformed into a form category. Not all data can be processed well by this technique. Data must be transformed into data category. In this technique, this is still done separately. Data transformation has not been conducted properly. The amount of cluster was not specified by the user, and the number of clusters generated may not match expectations. Likewise, big data variations and prevalent in every attribute will generate a lot of clusters, so it is difficult to distinguish from each other. This happens because the distance between the clusters have become very close. MDDS technique was tested on assessment data at the University, while experiments on elementary and secondary education data have not been conducted.

Bibliography

Ahmed, S.R.: Applications of data mining in retail business. Information Technology, Coding and Computing, 2, pp. 455–459 (2004).

Baepler, P. and Murdoch, C. J.: Academic Analitics and Data Mining in Higher Education. International Journal for the Scholarship of Teaching and Learning, 4(2), pp. 1–9 (2010).

Bagga, S., Singh, G.N.: Three Phase Iterative Model of KDD, International Journal of Information Technology and Knowledge Management, 4(2), pp. 695–697 (2011).

Blagojevic, M. and Micic, Z.: A web-based intelligent report e-learning system using data mining techniques. Computers and Electrical Engineering, pp. 1–10 (2012).
Deng, S., He, Z., and Xu, X.: G-ANMI: A mutual information based genetic clustering algorithm for categorical data. Knowledge-Based Systems, 23, pp. 144–149 (2012).
Giannotti, F., Gozzi, G., and Manco, G.: Clustering transactional data. In: Proceeding of PKDD'02, pp. 175–187 (2002).
Hongwu, Q., Xiuqin, M., Zain, J.M., and Herawan, T.: T. A Novel Soft Set Approach for Selecting Clustering Attribute. Knowledge-Based Systems, 36, pp. 139–145 (2012).
Hoppner, F. and Klawonn, F.: Learning Fuzzy Systems - An Objective Function-Approach. Mathware & Soft Computing, 11, pp. 143–162 (2004).
Han, J., Kamber, M., and Pei, J.: Data Mining Concept and Techniques. 3nd, Morgan kaufmann (2011).
Herawan, T.: Rough Clustering for Cancer Datasets. International Journal of Modern Physics: Conference Series, 9(1), pp. 240–258 (2012).
Kong Z., Zhang G., Wang L., Wu Z., Qi S., and Wang, H.: An efficient decision-making approach in incomplete soft set. Applied Mathematical Modelling, 38(7–8), pp. 2141–2150 (2014).
Kong, Z., Jia, W., Zhang, G., and Wang, L.: Normal parameter reduction in soft set based on particle swarm optimization algorithm. Applied Mathematical Modelling, 39, pp. 4808–4820 (2015).
Mamat, R., Herawan, T., and Deris, M.M.: MAR Maximum Attribute Relative of Soft Set-in selecting Clustering Attribute. To appear in Knowledge-Based Systems, pp. 1–10 (2013).
Maimon, O. and Rokach, L.: Data Mining and Knowledge Discovery Handbook 2nd ed., Springer Science+business media (2010).